Detecting Word Sense Disambiguation Biases in Machine Translation for Model-Agnostic Adversarial Attacks

Detecting Word Sense Disambiguation Biases in Machine Translation for Model-Agnostic Adversarial Attacks

Denis Emelin[1], Ivan Titov[1, 2], and Rico Sennrich[3, 1]

[1]University of Edinburgh, Scotland
[2]University of Amsterdam, Netherlands
[3]University of Zurich, Switzerland
D.Emelin@sms.ed.ac.uk
ititov@inf.ed.ac.uk sennrich@cl.uzh.ch

Abstract

Word sense disambiguation is a well-known source of translation errors in NMT. We posit that some of the incorrect disambiguation choices are due to models' over-reliance on dataset artifacts found in training data, specifically superficial word co-occurrences, rather than a deeper understanding of the source text. We introduce a method for the prediction of disambiguation errors based on statistical data properties, demonstrating its effectiveness across several domains and model types. Moreover, we develop a simple adversarial attack strategy that minimally perturbs sentences in order to elicit disambiguation errors to further probe the robustness of translation models. Our findings indicate that disambiguation robustness varies substantially between domains and that different models trained on the same data are vulnerable to different attacks.[1]

1 Introduction

Consider the sentence *John met his wife in the hot spring of 1988*. In this context, the polysemous term *spring* unambiguously refers to the season of a specific year. Its appropriate translation into German would therefore be *Frühling* (the season), rather than one of its alternative senses, such as *Quelle* (the source of a stream). To contemporary machine translation systems, however, this sentence presents a non-trivial challenge, with Google Translate (GT) producing the following translation: *John traf seine Frau in der heißen Quelle von 1988.*

Prior studies have indicated that neural machine translation (NMT) models rely heavily on source sentence information when resolving lexical ambiguity (Tang et al., 2019). This suggests that the combined source contexts in which a specific sense of an ambiguous term occurs in the training data greatly inform the models' disambiguation decisions. Thus, a stronger correlation between the English collocation *hot spring* and the German translation *Quelle*, as opposed to *Frühling*, in the training corpus may explain this disambiguation error. Indeed, *John met his wife in the spring of 1988* is translated correctly by GT.

We propose that our motivating example is representative of a systematic pathology NMT systems have yet to overcome when performing word sense disambiguation (WSD). Specifically, we hypothesize that translation models learn to disproportionately rely on lexical correlations observed in the training data when resolving word sense ambiguity. As a result, disambiguation errors are likely to arise when an ambiguous word co-occurs with words that are strongly correlated in the training corpus with a sense that differs from the reference.

To test our hypothesis, we evaluate whether dataset artifacts are predictive of disambiguation decisions made in NMT. First, given an ambiguous term, we define a strategy for quantifying how much its context biases NMT models towards its different target senses, based on statistical patterns in the training data. We validate our approach by examining correlations between this bias measure and WSD errors made by baseline models. Furthermore, we investigate whether such biases can be exploited for the generation of minimally-perturbed adversarial samples that trigger disambiguation errors. Our method does not require access to gradient information nor the score distribution of the decoder, generates samples that do not significantly diverge from the training domain, and comes with a clearly-defined notion of attack success and failure.

The main contributions of this study are:

1. We present evidence for the over-reliance of NMT systems on inappropriate lexical correlations when translating polysemous words.

[1]Experimental codebase available at http://github.com/demelin/detecting_wsd_biases_for_nmt

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 7635–7653,
November 16–20, 2020.

2. We propose a method for quantifying WSD biases that can predict disambiguation errors.

3. We leverage data artifacts for the creation of adversarial samples that facilitate WSD errors.

2 Can WSD errors be predicted?

To evaluate whether WSD errors can be effectively predicted, we first propose a method for measuring the bias of sentence contexts towards different senses of polysemous words, based on lexical co-occurrence statistics of the training distribution. We restrict our investigation to English→German, although the presented findings can be assumed to be language-agnostic. To bolster the robustness of our results, we conduct experiments in two domains - movie subtitles characterized by casual language use, and the more formal news domain. For the former, we use the OpenSubtitles2018 (OS18) (Lison et al., 2019) corpus[2], whereas the latter is represented by data made available for the news translation task of the Fourth Conference on Machine Translation (WMT19)[3] (Barrault et al., 2019). Appendix A.1 reports detailed corpus statistics.

2.1 Quantifying disambiguation biases

An evaluation of cross-lingual WSD errors presupposes the availability of certain resources, including a list of ambiguous words, a lexicon containing their possible translations, and a set of parallel sentences serving as a disambiguation benchmark.

Resource collection

Since lexical ambiguity is a pervasive feature of natural language, we limit our study to homographs - polysemous words that share their written form but have multiple, unrelated meanings. We further restrict the set of English homographs to nouns that are translated as distinct German nouns, so as to confidently identify disambiguation errors, while minimizing the models' ability to disambiguate based on syntactic cues. English homographs are collected from web resources[4], excluding those that do not satisfy the above criteria. Refer to appendix A.2 for the full homograph list.

We next compile a parallel lexicon of homograph translations, prioritizing a high coverage of all possible senses. Similar to (Raganato et al., 2019), we obtain sense-specific translations from cross-lingual BabelNet (Navigli and Ponzetto, 2010) synsets. Since BabelNet entries vary in their granularity, we iteratively merge related synsets as long as they have at least three German translations in common or share at least one definition.[5] This leaves us with multiple sense clusters of semantically related German translations per homograph. To further improve the quality of the lexicon, we manually clean and extend each homograph entry to address the noise inherent in BabelNet and its incomplete coverage.[6] Appendix A.7 provides examples of the final sense clusters.

In order to identify sentence contexts specific to each homograph sense, parallel sentences containing known homographs are extracted from the training corpora in both domains. We lemmatize homographs, their senses, and all sentence pairs using spaCy (Honnibal and Montani, 2017) to improve the extraction recall. Homographs are further required to be aligned with their target senses according to alignments learned with fast_align (Dyer et al., 2013). Each extracted pair is assigned to one homograph sense cluster based on its reference homograph translation. Pairs containing homograph senses assigned to multiple clusters are ignored, as disambiguation errors are impossible to detect in such cases.

Bias measures

It can be reasonably assumed that context words co-occurring with homographs in a corpus of natural text are more strongly associated with some of their senses than others. Words that are strongly correlated with a specific sense may therefore bias models towards the corresponding translation at test time. We refer to any source word that co-occurs with a homograph as an *attractor* associated with the sense cluster of the homograph's translation. Similarly, we denote the degree of an attractor's association with a sense cluster as its *disambiguation bias* towards that cluster. Table 1 lists the most frequent attractors identified for the different senses of the homograph *spring* in the OS18 training set.

Intuitively, if an NMT model disproportionately relies on simple surface-level correlations when resolving lexical ambiguity, it is more likely to make WSD errors when translating sentences that contain

[2] http://opus.nlpl.eu

[3] http://statmt.org/wmt19

[4] http://7esl.com/homographs http://en.wikipedia.org/wiki/List_of_English_homographs

[5] A manual inspection found the clusters to be meaningful.

[6] The lexicon is released as part of our experimental code: http://github.com/demelin/detecting_wsd_biases_for_nmt.

season	*water source*	*device*
summer	hot	like
winter	water	back
come	find	thing

Table 1: Examples of attractors for *spring*.

strong attractors towards a wrong sense. To test this, we collect attractors from the extracted parallel sentences, quantifying their disambiguation bias (DB) using two metrics: Raw co-occurrence frequency (FREQ) and positive point-wise mutual information (PPMI) between attractors and homograph senses. FREQ is defined in Eqn.1, while Eqn.2 describes PPMI, with $w \in V$ denoting an attractor term in the source vocabulary[7], and $sc \in SC$ denoting a sense cluster in the set of sense clusters assigned to a homograph. For PPMI, $P(w_i, sc_j)$, $P(w_i)$, and $P(sc_j)$ are estimated via relative frequencies of (co-)occurrences in training pairs.

$$FREQ(w_i, sc_j) = Count(w_i, sc_j) \quad (1)$$

$$PPMI(w_i, sc_j) = max(\frac{P(w_i, sc_j)}{P(w_i)P(sc_j)}, 0) \quad (2)$$

The disambiguation bias associated with the entire context of a homograph is obtained by averaging sense-specific bias values $DB(w_i, sc_j)$ of all attractors in the source sentence $S = \{w_1, w_2, ..., w_{|S|}\}$, as formalized in Eqn.3. Context words that are not known attractors of sc_j are assigned a disambiguation bias value of 0.

$$DB(S, sc_j) = \frac{1}{|S|} \sum_{i=1}^{|S|} DB(w_i, sc_j) \quad (3)$$

As a result, sentences containing a greater number of strong attractors are assigned a higher bias score.

2.2 Probing NMT models

To evaluate the extent to which sentence-level disambiguation bias is predictive of WSD errors made by NMT systems, we train baseline translation models for both domains. The baselines include Transformer (Vaswani et al., 2017), LSTM (Luong et al., 2015), and convolutional Seq-to-Seq (ConvS2S) (Gehring et al., 2017) models of comparable size. Appendix A.4 details the training regime and hyper-parameter choices. SacreBLEU (Post, 2018) scores reported in Table 2 indicate that all models are reasonably competent.

		WMT	
Architecture	**OS18 test**	**test 2014**	**test 2019**
Transformer	29.7	27.5	38.2
LSTM	27.7	24.1	34.3
ConvS2S	27.7	23.5	32.5

Table 2: EN-DE translation performance (BLEU).

Test sets for WSD error prediction are constructed by extracting parallel sentences from held-out, development, and test data (see appendix A.1 for details). The process is identical to that described in section 2.1, with the added exclusion of source sentences shorter than 10 tokens, as they may not provide enough context. For each source sentence, disambiguation bias values are computed according to equation 3, with sc_j corresponding to either the correct sense cluster ($DB_{\checkmark}$) or the incorrect sense cluster with the strongest bias ($DB_{\times}$). Additionally, we consider the difference DB_{DIFF} between $DB_{\times}$ and $DB_{\checkmark}$ which can be interpreted as the overall statistical bias in a source sentence towards an incorrect homograph translation. All bias scores are computed either using FREQ or PPMI.

We examine correlations between the proposed bias measures and WSD errors produced by the in-domain baseline models. Translations are considered to contain WSD errors if the target homograph sense does not belong to the same sense cluster as its reference translation. We check this by looking up target words aligned with source homographs according to fast_align. To estimate correlation strength we employ the ranked biserial correlation (RBC) metric[8] (Cureton, 1956) and measure statistical significance using the Mann-Whitney U (MWU) test (Mann and Whitney, 1947).

In order to compute the RBC values, test sentences are divided into two groups - one containing correctly translated source sentences and another comprised of source sentences with incorrect homograph translations. Next, all possible pairs are constructed between the two groups, pairing together each source sentence from one group with all source sentences from the other. Finally, the

[7] We consider any word that co-occurs with a homograph in the training corpus as an attractor of the homograph's specific sense cluster, except for the homograph itself which is not regarded as an attractor for any of its known sense clusters.

[8] We additionally used the non-parametric generalization of the Common Language Effect Size (Ruscio, 2008) for correlation size estimation, but couldn't detect any advantages over RBC in our experimental setting.

Model	FREQ✓	PPMI✓	FREQ✗	PPMI✗	$\text{FREQ}_{\text{DIFF}}$	$\text{PPMI}_{\text{DIFF}}$	Length
OS18 Transformer	-0.532	-0.578	0.327	0.474	**0.708**	0.674	0.018
OS18 LSTM	-0.468	-0.504	0.386	0.486	**0.690**	0.630	0.008
OS18 ConvS2S	-0.477	-0.514	0.391	0.492	**0.723**	0.658	0.021
WMT19 Transformer	-0.610	-0.668	0.415	0.579	**0.687**	0.677	-0.004
WMT19 LSTM	-0.661	-0.698	0.376	0.574	**0.725**	0.708	-0.009
WMT19 ConvS2S	-0.648	-0.678	0.408	0.599	**0.731**	0.710	0.000

Table 3: Rank biserial correlation between disambiguation bias measures and lexical disambiguation errors.

proportion of pairs f where the DB score of the incorrectly translated sentence is greater than that of the correctly translated sentence is computed, as well as the proportion of pairs u where the opposite relation holds. The RBC value is then obtained according to Eqn.4.

$$RBC = f - u \tag{4}$$

Statistical significance, on the other hand, is estimated by ranking all sentences in the test set according to their DB score in ascending order while resolving ties, and computing the U-value according to Eqn.5-7, where R_1 denotes to the sum of ranks of sentences with incorrectly translated homographs and n_1 their total count, while R_2 denotes the sum of ranks of correctly translated sentences and n_2 their respective total count.

$$U = min(U_1, U_2) \tag{5}$$

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2} \tag{6}$$

$$U_2 = R_2 - \frac{n_2(n_2+1)}{2} \tag{7}$$

To obtain the p-values, U-values are subjected to tie correction and normal approximation.[9]

Table 3 summarizes the results[10], including correlation estimates between WSD errors and source sentence length, as a proxy for disambiguation context size. Statistically significant correlations are discovered for all bias estimates based on attractors ($p < 1e\text{-}5$, two-sided). Moreover, the observed correlations exhibit a strong effect size (McGrath and Meyer, 2006). See appendix A.5 for the model-specific effect size interpretation thresholds. For all models and domains the strongest correlations are observed for DB_{DIFF} derived from simple co-occurrence counts.

[9]We use Python implementations of RBC and MWU provided by the pingouin library (Vallat, 2018).

[10]Positive values denote a positive correlation between bias measures and the presence of disambiguation errors in model translations, whereas negative values denote negative correlations. The magnitude of the values, meanwhile, indicates the correlations' effect size.

Challenge set evaluation

To establish the predictive power of the uncovered correlations, a challenge set of 3000 test pairs with the highest $\text{FREQ}_{\text{DIFF}}$ score is subsampled from the full WSD test pair pool in both domains. In addition, we create secondary sets of equal size by randomly selecting pairs from each pool. As Figure 1 illustrates, our translation models exhibit a significantly higher WSD error rate - by a factor of up to **6.1** - on the challenge sets as compared to the randomly chosen pairs. While WSD performance is up to 96% on randomly chosen sentences, performance drops to 77–82% for the best-performing model (Transformer). This suggests that lexical association artifacts, from which the proposed disambiguation bias measure is derived, can be an effective predictor of lexical ambiguity resolution errors across model architectures and domains.

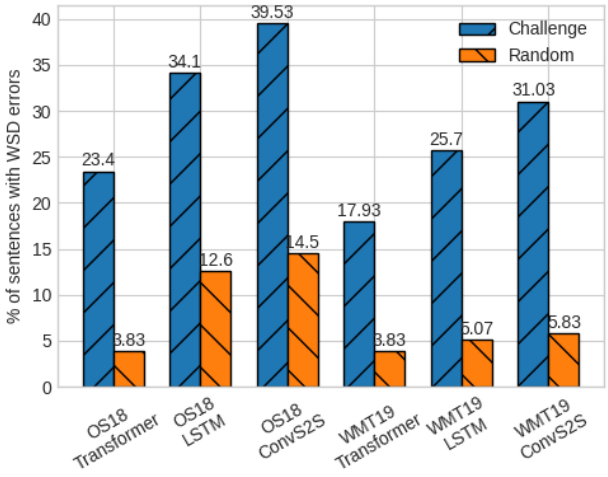

Figure 1: WSD errors in subsampled challenge sets.

The observed efficacy of attractor co-occurrence counts for WSD error prediction may be partially due to sense frequency effects, since more frequent senses occur in more sentence pairs, yielding more frequent attractors. NMT models are known to underperform on low-frequency senses of ambiguous terms (Rios et al., 2017), prompting us to investigate if disambiguation biases capture the same information. For this purpose, another challenge set of 3000 pairs is constructed by prioritizing pairs assigned to the rarest among each homograph's sense sets. We find that the new challenge set has a 72.63% overlap with the disambiguation bias challenge set in the OS18 domain and 64.4% overlap in the WMT19 domain. Thus, disambiguation biases appear to indeed capture some sense frequency effects, which themselves represent a dataset artifact, but also introduce novel information.

Our experimental findings indicate that translation models leverage undesirable surface-level correlations when resolving lexical ambiguity and are prone to disambiguation errors in cases where learned statistical patterns are violated. Next, we use these insights for the construction of adversarial samples that cause disambiguation errors by minimally perturbing source sentences.

3 Adversarial WSD attacks on NMT

Adversarial attacks probe model robustness by attempting to elicit incorrect predictions with perturbed inputs (Zhang et al., 2020). By crafting adversarial samples that explicitly target WSD capabilities of NMT models, we seek to provide further evidence for their susceptibility to dataset artifacts.

3.1 Generating adversarial WSD samples

Our proposed attack strategy is based on the assumption that introducing an attractor into a sentence can flip its inherent disambiguation bias towards the attractor's sense cluster. Thus, translations of the so perturbed sentence will be more likely to contain WSD errors. The corresponding sample generation strategy consists of four stages:

1. Select *seed* sentences containing homographs to be adversarially perturbed.

2. Identify attractors that are likely to yield fluent and natural samples.

3. Apply perturbations by introducing attractors into seed sentences.

4. Predict effective adversarial samples based on attractor properties.

The targeted attack is deemed successful if a victim model accurately translates the homograph in the seed sentence, but fails to correctly disambiguate it in the adversarially perturbed sample, instead translating it as one of the senses belonging to the attractor's sense cluster. This is a significantly more challenging attack success criterion than the general reduction in test BLEU typically employed for evaluating adversarial attacks on NMT systems (Cheng et al., 2019). Samples are generated using homographs and attractors collected in section 2.1, while all test sentence pairs extracted in section 2.2 form the domain-specific seed sentence pools. Attack success is evaluated on the same baseline translation models as used throughout section 2.

Seed sentence selection

In order to generate informative and interesting adversarial samples, we focus on seed sentences that are likely to be unambiguous. We thus apply three filtering heuristics to seed sentence pairs:

- Sentences have to be at least 10 tokens long.

- We mask out the correct homograph sense in the reference translation and use a pre-trained German BERT model (Devlin et al., 2019)[11] to predict it. Pairs are rejected if the most probable sense does not belong to the correct sense cluster which suggests that the sentence context may be insufficient for correctly disambiguating the homograph. As a result, WSD errors observed in model-generated translations of the constructed adversarial samples are more likely to be due to the applied adversarial perturbations.

- 10% of pairs with the highest disambiguation bias towards incorrect sense clusters are removed from the seed pool.

Setting the rejection threshold above 10% can further reduce WSD errors in seed sentences. At the same time, it would likely render minimal perturbations ineffective, due to the sentences' strong bias towards the correct homograph sense. Thus, we aim for a working compromise.

[11] We use the implementation provided by the Hugging Face Transformers library (Wolf et al., 2019). We do not fine-tune BERT, as our use case corresponds to its original masked language modeling objective.

IH	During this **first** **spring**, he planted another tree that looked the same.
RH	A **hot** ~~new~~ **spring** will conquer the dark nights of winter.
InH	Come the **spring**, I will be invading the **whole** country called Frankia.
RnH	After a long, **eternal** ~~fallow~~ winter, **spring** has come again to Fredericks Manor.

Table 4: Perturbation examples; seed sense: *season*, adversarial sense: *water source*. Insertion/replacement in red.

Perturbation types

Naively introducing new words into sentences is expected to yield disfluent, unnatural samples. To counteract this, we constrain candidate attractors to adjectives, since they can usually be placed in front of English nouns without violating grammatical constraints. We consider four perturbation types:

- Insertion of the attractor adjective in front of the homograph (IH)
- Replacement of a seed adjective modifying the homograph (RH)
- Insertion of the attractor adjective in front of a non-homograph noun (InH)
- Replacement of a seed adjective modifying a non-homograph noun (RnH)

Replacement strategies require seed sentences to contain adjectives, but can potentially have a greater impact on the sentence's disambiguation bias by replacing attractors belonging to the correct sense cluster. Examples for each generation strategy are given in Table 4, with homographs highlighted in blue and added attractors in red.

Attractor selection

Since adjectives are subject to selectional preferences of homograph senses, not every attractor will yield a semantically coherent adversarial sample. For instance, inserting the attractor *flying* in front of the homograph *bat* in a sentence about baseball will likely produce a nonsensical expression, whereas an attractor like *huge* would be more acceptable. We attempt to control for this type of disfluency by only considering attractors that had been previously observed to modify the homograph in its seed sentence sense. For non-homograph perturbations, attractors must have been observed modifying the non-homograph noun. This is ensured by obtaining a dependency parse for each sentence in the English half of the training data and maintaining a list of modifier adjectives for each known target homograph sense set and source noun.[12]

Lastly, to facilitate the fluency and naturalness of adversarial samples, the generation process incorporates a series of constraints:

- Comparative and superlative adjective forms are excluded from the attractor pool.
- Attractors may not modify compound nouns due to less transparent selectional preferences.
- Attractors are not allowed next to other adjectives modifying the noun, to avoid violating the canonical English adjective order.

As all heuristics rely on POS taggers or dependency parsers,[13] they are not free of noise, occasionally yielding disfluent or unnatural samples.

We restrict the number of insertions or replacements to one, so as to maintain a high degree of semantic similarity between adversarial samples and seed sentences. A single seed sentence usually yields several samples, even after applying the aforementioned constraints. Importantly, we generate samples using all retained attractors at this stage, without selecting for expected attack success.

Post-generation filtering

To further ensure the naturalness of generated samples, sentence-level perplexity is computed for each seed sentence and adversarial sample using a pretrained English GPT2 (Radford et al., 2019) language model.[14] Samples are rejected if their perplexity exceeds that of their corresponding seed sentence by more than 20%. In total, we obtain a pool of ∼500K samples for the OS18 domain and ∼3.9M samples for the WMT19 domain. Each sample is translated by all in-domain models.

3.2 Identifying effective attractors

The success of the proposed attack strategy relies on the selection of attractors that are highly likely

[12] This assumes correctness of homograph reference translations, which is unfortunately not always guaranteed.

[13] We use spaCy in all cases.

[14] As implemented in the Transformers library.

Model	$\text{FREQ}_{✗}$	$\text{PPMI}_{✗}$	$\text{FREQ}_{\text{DIFF}}$	$\text{PPMI}_{\text{DIFF}}$
OS18 Transformer	0.307	0.367	**0.438**	0.306
OS18 LSTM	0.258	0.261	**0.375**	0.227
OS18 ConvS2S	0.228	0.174	**0.325**	0.165
WMT19 Transformer	0.241	0.241	**0.264**	0.224
WMT19 LSTM	0.278	0.256	**0.316**	0.231
WMT19 ConvS2S	0.304	0.270	**0.328**	0.216

Table 5: Rank biserial correlation between attractors' disambiguation bias and attack success.

to flip the homograph translation from the correct *seed* sense towards an *adversarial* sense belonging to the attractors' own sense set. To identify such attractors, we examine correlations between attractors' disambiguation biases and the effectiveness of adversarial samples containing them. The attractors' bias values are based either on co-occurrence frequencies (Eqn.1) or PPMI scores (Eqn.2) with the homographs' sense clusters. In particular, we examine the predictive power of an attractor's bias towards the adversarial sense cluster ($\text{DB}_{✗}$) as well as the difference between its adversarial and seed bias values (DB_{DIFF}). As before, RBC and MWU measures are used to estimate correlation strength, with Table 5 summarizing the results.

Similarly to the findings reported in section 2.2, all uncovered correlations are strong and statistically significant with $p < 1e\text{-}5$ (see appendix A.5 for effect size thresholds). Importantly, $\text{FREQ}_{\text{DIFF}}$ exhibits the strongest correlation in all cases.

We are furthermore interested in establishing which of the proposed perturbation methods yields most effective attacks. For this purpose, we examine the percentage of attack successes per perturbation strategy in Figure 2, finding perturbations proximate to the homograph to be most effective.

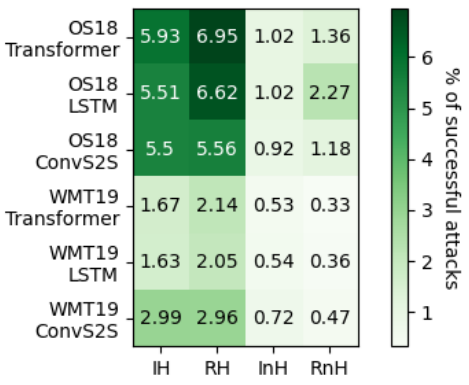

Figure 2: Successful attacks per perturbation.

Challenge set evaluation

Having thus identified a strategy for selecting attractors that are likely to yield successful attacks, we construct a challenge set of 10000 adversarial samples with the highest attractor $\text{FREQ}_{\text{DIFF}}$ scores that had been obtained via the IH or RH perturbations. To enforce sample diversity, we limit the number of samples to at most 1000 per homograph. Additionally, we create equally-sized, secondary challenge sets by drawing samples at random from each domain's sample pool. Figure 3 illustrates the attack success rate for both categories, while Table 6 shows some of the successful attacks on the OS18 transformer. Further successful samples are reported in Appendix A.7.

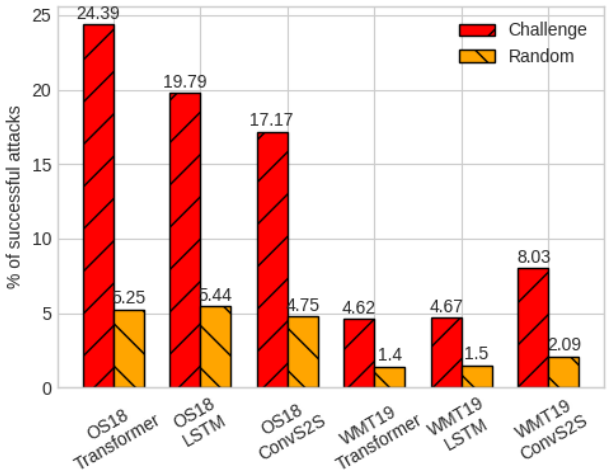

Figure 3: Successful challenge sets attacks.

The success rates are modest, ranging from 4.62% to 24.39%, but nonetheless showcase the capacity of targeted, minimal perturbations for flipping correct homograph translations towards a specific sense set. Since our attacks do not require access to model gradients or predictive score distributions, fall within the same domain as the models' training data, and have a strict notion of success,

Source input / Original output / Perturbed output	Seed sense	Adv. sense
S: We played the songs again until we felt they sounded right, worked out all the (**nasty**) **bugs**. **O**: Wir spielten die Lieder wieder, bis sie sich richtig anhörten und alle **Fehler**✓ ausarbeiteten. **P**: Wir spielten die Lieder wieder, bis sie sich richtig anhörten und alle **bösen Käfer**✗ ausarbeiteten.	*error*	*insect*
S: The driver gets out, opens the (**large**) **boot**, takes some flowers out to deliver. **O**: Der Fahrer steigt aus, öffnet den **Kofferraum**✓, nimmt ein paar Blumen zum Ausliefern mit. **P**: Der Fahrer steigt aus, öffnet den **großen Stiefel**✗, nimmt ein paar Blumen zum Ausliefern mit.	*trunk*	*shoe*
S: The doctor somehow got that wig mixed up with the newspapers and (**different**) **letters**. **O**: Der Arzt verwechselte die Perücke mit den Zeitungen und **Briefen**✓. **P**: Der Arzt verwechselte die Perücke mit den Zeitungen und **anderen Buchstaben**✗.	*message*	*character*
S: And he will not cease until every **last race** of the Four Lands is destroyed. **O**: Und er wird nicht aufgeben, bis jede **Rasse**✓ der Vier Länder ausgelöscht ist. **P**: Und er wird nicht aufhören, bis jedes **letzte Rennen**✗ der Vier Länder zerstört ist.	*ethnic group*	*contest*

Table 6: Examples of successful attacks on the OS18 transformer. Homographs are blue, attractors are red.

direct comparisons with previous work are difficult. Crucially, compared with a random sample selection strategy, subsampling informed by attractors' disambiguation bias is up to **4.25** times more successful at identifying effective adversarial samples.

While the relative improvement in attack success rate over the random baseline is comparable in both domains, the OS18 models are more susceptible to attacks in absolute terms. This may be due to their lower quality, or the properties of the training data, which can suffer from noisiness (Lison et al., 2019). Interestingly, the relative robustness of individual model architectures to WSD attacks also differs between domains, despite similar quality in terms of BLEU (see Table 2). A more thorough investigation of architecture-specific WSD vulnerabilities is left for future work.

3.3 Sample quality analysis

To examine whether our adversarial samples would appear trivial and innocuous to human translators, automatic and human evaluation of samples included in the challenge set is conducted. Following (Morris et al., 2020), we use a grammar checker[15] to evaluate the number of cases in which adversarial perturbations introduce grammatical errors. In the OS18 domain, only 1.04% of samples are less grammatical than their respective seed sentences, whereas this is the case for 2.04% of WMT19 samples, indicating a minimal degradation.

We additionally present two bilingual judges with 1000 samples picked at random from adversarial challenge sets in both domains and 1000 regular sentences from challenge sets constructed in section 2.2. For each adversarial source sentence, annotators were asked to choose whether the homograph's translation belongs to the correct or adversarial seed cluster. For each regular sentence, the choice was between the correct and randomly selected clusters. Across both domains, annotator error rate was 11.23% in the adversarial setting and 11.45% for regular sentences. As such, the generated samples display a similar degree of ambiguity to natural sentences that are likely to elicit WSD errors in NMT models. Annotator agreement was substantial (Cohen's kappa = 0.7).

The same judges were also asked to rate the naturalness of each sentence on a Likert scale from 1 to 5. Perturbed sentences were assigned a mean score of 3.94, whereas regular sentences scored higher at 4.18. However, annotator agreement was low (weighted Kappa = 0.17). The observed drop in naturalness is likely due to the selection of attractors that are not fully consistent with the selectional preferences of homograph senses during sample generation. We attribute this to WSD errors in reference translations. For instance, we find that the attractor *vampire* is occasionally applied to seed sentences containing the homograph *bat* in its *sporting equipment* sense, which can only occur if the attractor has been observed to modify this sense cluster in the training data (see 3.1). Appendix A.6 replicates annotator instructions for both tasks.

4 Transferability of adversarial samples

An interesting question to consider is whether translation models trained on the same data are vulnerable to the same adversarial samples. We evaluate this by computing the Jaccard similarity index between successful attacks on each baseline model from the entire pool of adversarial samples

[15] http://languagetool.org

described in section 3.2. We find the similarity to be low, raging between 10.1% and 18.2% for OS18 and between 5.7% and 9.1% for WMT19 samples, which suggests that different model architectures appear to be sensitive to different corpus artifacts, possibly due to differences in their inductive biases.

Considering the observed discrepancy in vulnerabilities between architectures, a natural follow-up question is whether two different instances of the same architecture are susceptible to the same set of attacks. We investigate this by training a second transformer model for each domain, keeping all settings constant with the initial models, but choosing a different seed for the random initialization. While the similarity between sets of successful adversarial samples is greater for two models of the same type, with 25.2% in the OS18 and 12.4% in WMT19 domain, is it still remarkably low.

5 Literature review

Polysemous terms represent a long-standing challenge for NMT. Past investigations sought to quantify the WSD capacity of translation models through challenge sets (Rios et al., 2017; Raganato et al., 2019), to understand the disambiguation process by analysing their internal representations (Marvin and Koehn, 2018; Tang et al., 2019), or to improve ambiguity resolution capabilities of translation models (Rios et al., 2017; Liu et al., 2018). To our knowledge, no study so far has examined the interaction between training data artifacts and WSD performance in detail.

Dataset artifacts, on the other hand, have previously been shown to enable models to make correct predictions based on incorrect or insufficient information (McCoy et al., 2019; Gururangan et al., 2018) by over-relying on spurious correlations present in the training data. Within NMT, models were found to exhibit gender-bias, reinforcing harmful stereotypes (Vanmassenhove et al., 2018; Stanovsky et al., 2019). As a response, strategies have been proposed for de-biasing the training data (Li and Vasconcelos, 2019; Le Bras et al., 2020), as well as for making models more robust to data biases through adversarial training (Belinkov et al., 2019).

Adversarial attacks have recently been extended as an effective model analysis tool from vision to language tasks (Samanta and Mehta, 2017; Alzantot et al., 2018; Glockner et al., 2018; Zhang et al., 2019), including NMT (Cheng et al., 2018, 2019), where the focus so far has been on strategies requiring direct access to the victim model's loss gradient or output distribution. Recent surveys suggested that state-of-the-art attacks often yield ungrammatical and meaning-destroying samples, thus diminishing their usefulness for the evaluation of model robustness (Michel et al., 2019; Morris et al., 2020). Targeted attacks on WSD abilities of translation models have so far remained unexplored.

6 Conclusion

We conducted an initial investigation into leveraging data artifacts for the prediction of WSD errors in machine translation and proposed a simple adversarial attack strategy based on the presented insights. Our results show that WSD is not yet a solved problem in NMT, and while the general performance of popular model architectures is high, we can identify or create sentences where models are more likely to fail due to data biases.

The effectiveness of our methods owes to neural models struggling to accurately distinguish between meaningful lexical correlations and superficial ones. As such, the presented approach is expected to be transferable to other language pairs and translation directions, assuming that the employed translation models share this underlying weakness. Given the model-agnostic nature of our findings, this is likely to be the case.

As a continuation to this work, we intend to evaluate whether multilingual translation models are more resilient to lexical disambiguation biases and, as a consequence, are less susceptible to adversarial attacks that exploit source-side homography. Extending model-agnostic attack strategies to incorporate other types of dataset biases and to target natural language processing tasks other than machine translation is likewise a promising avenue for future research. Lastly, the targeted development of models that are resistant to dataset artifacts is a promising direction that is likely to aid generalization across linguistically diverse domains.

Acknowledgements

We thank Sabine Weber and Tom Pelsmaeker for valuable discussions throughout the development of this work, as well as the anonymous reviewers for their constructive feedback. Rico Sennrich has received funding from the Swiss National Science Foundation (project MUTAMUR; no. 176727).

References

Moustafa Alzantot, Yash Sharma Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Loïc Barrault, Ondřej Bojar, Marta R Costa-Jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. 2019. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61.

Yonatan Belinkov, Adam Poliak, Stuart Shieber, Benjamin Van Durme, and Alexander Sasha Rush. 2019. On adversarial removal of hypothesis-only bias in natural language inference. In *Proceedings of the Joint Conference on Lexical and Computational Semantics*.

Minhao Cheng, Jinfeng Yi, Huan Zhang, Pin-Yu Chen, and Cho-Jui Hsieh. 2018. Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples. *arXiv preprint arXiv:1803.01128*.

Yong Cheng, Lu Jiang, and Wolfgang Macherey. 2019. Robust neural machine translation with doubly adversarial inputs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4324–4333.

Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. Academic press.

Edward E Cureton. 1956. Rank-biserial correlation. *Psychometrika*, 21(3):287–290.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, and Yann Dauphin. 2017. A convolutional encoder model for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 123–135.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking nli systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1).

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. *arXiv*, pages arXiv–2002.

Yi Li and Nuno Vasconcelos. 2019. Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9572–9581.

Pierre Lison, Jörg Tiedemann, Milen Kouylekov, et al. 2019. Open subtitles 2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *LREC 2018, Eleventh International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA).

Frederick Liu, Han Lu, and Graham Neubig. 2018. Handling homographs in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1336–1345.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

Rebecca Marvin and Philipp Koehn. 2018. Exploring word sense disambiguation abilities of neural machine translation systems (non-archival extended abstract). In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 125–131.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448.

Robert E McGrath and Gregory J Meyer. 2006. When effect sizes disagree: the case of r and d. *Psychological methods*, 11(4):386.

Paul Michel, Xian Li, Graham Neubig, and Juan Miguel Pino. 2019. On evaluation of adversarial perturbations for sequence-to-sequence models. In *Proceedings of NAACL-HLT*, pages 3103–3114.

John X Morris, Eli Lifland, Jack Lanchantin, Yangfeng Ji, and Yanjun Qi. 2020. Reevaluating adversarial examples in natural language. *arXiv preprint arXiv:2004.14174*.

Roberto Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. 2019. The mucow test suite at wmt 2019: Automatically harvested multilingual contrastive word sense disambiguation test sets for machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 470–480.

Annette Rios, Laura Mascarell, and Rico Sennrich. 2017. Improving word sense disambiguation in neural machine translation with sense embeddings. In *Proceedings of the Second Conference on Machine Translation*, pages 11–19.

John Ruscio. 2008. A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological methods*, 13(1):19.

Suranjana Samanta and Sameep Mehta. 2017. Towards crafting text adversarial samples. *arXiv preprint arXiv:1707.02812*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Gabriel Stanovsky, Noah A Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684.

Gongbo Tang, Rico Sennrich, and Joakim Nivre. 2019. Encoders help you disambiguate word senses in neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1429–1435.

Raphael Vallat. 2018. Pingouin: statistics in python. *The Journal of Open Source Software*, 3(31):1026.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Huangzhao Zhang, Hao Zhou, Ning Miao, and Lei Li. 2019. Generating fluent adversarial examples for natural languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5569.

Wei Emma Zhang, Quan Z Sheng, Ahoud Alhazmi, and Chenliang Li. 2020. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3):1–41.

A Supplementary material

A.1 Data properties

The WMT19 data is obtained by concatenating the Europarl v9, Common Crawl, and News Commentary v14 parallel corpora. Basic data cleaning is performed for both domains, which includes removal of pairs containing sentences classified by langid[16] as neither German or English and pairs with a source-to-target sentence length ratio exceeding 2. We create development and training splits for the OS18 domain by removing 10K sentence pairs from the full, shuffled corpus in each case. For each domain, we additionally hold-out 20% of pairs to be used for the extraction of test pairs containing homographs, as described in section 2.2. Final statistics for the OS18 domain are reported in table 9 and in 10 for the WMT19 domain.

Each dataset is subsequently tokenized and truecased using Moses (Koehn et al., 2007) scripts[17]. For model training and evaluation, we additionally learn and apply BPE codes (Sennrich et al., 2016) to the data using the subword-NMT implementation[18], with 32k merge operations and the vocabulary threshold set to 50.

A.2 Homograph list

The full list of homographs used in our experiments is as follows: anchor, arm, band, bank, balance, bar, barrel, bark, bass, bat, battery, beam, board, bolt, boot, bow, brace, break, bug, butt, cabinet, capital, case, cast, chair, change, charge, chest, chip, clip, club, cock, counter, crane, cycle, date, deck, drill, drop, fall, fan, file, film, flat, fly, gum, hoe, hood, jam, jumper, lap, lead, letter, lock, mail, match, mine, mint, mold, mole, mortar, move, nail, note, offense, organ, pack, palm, pick, pitch, pitcher, plaster, plate, plot, pot, present, punch, quarter, race, racket, record, ruler, seal, sewer, scale, snare, spirit, spot, spring, staff, stock, subject, tank, tear, term, tie, toast, trunk, tube, vacuum, watch.

A.3 Sense cluster examples

Table 11 lists some of the identified sense clusters for several homographs. All homographs used in our experiments have at least two sense clusters associated with them.

[16] http://github.com/saffsd/langid.py

[17] http://github.com/moses-smt/mosesdecoder

[18] http://github.com/rsennrich/subword-nmt

A.4 Baseline models

Table 12 provides implementation and training details for each architecture. Same settings are used for training identical models types in different domains. We use standard fairseq[19] (Ott et al., 2019) implementations for all model types and train them on NVIDIA 1080ti or NVIDIA 2080ti GPUs. Model translations are obtained by averaging the final 5 model checkpoints and decoding using beam search with beam size 5.

A.5 Base-rate adjusted effect size thresholds

Whether the effect size of correlations between dichotomous and quantitative variables can be considered strong depends on the size ratio between the two groups denoted by the dichotomous variable, i.e. its base rate. As the standard formulation of RBC is sensitive to the base rate, the estimated effect size decreases as the base rate becomes more extreme (see (McGrath and Meyer, 2006) for details). Applied to our experimental setting, this means that the observed correlation values are sensitive to the number of sentences containing disambiguation errors relative to the amount of those that do not. This is an undesirable property, as we are only interested in the predictive power of our quantitative variables, regardless of how often disambiguation errors are observed. Thus, we adjust the thresholds for the interpretation of correlation strength to account for WSD errors being less frequent than WSD successes overall, in analogy to (McGrath and Meyer, 2006). Doing so enables the direct comparison of correlation strength between domains and model types, as each combination of the two factors exhibits a different disambiguation success base rate.

A common practice for interpreting effect size strength that does not account for base rate inequalities is the adoption of Cohen's benchmark (Cohen, 2013), which posits that the effect size d is large if $d >= 0.8$, medium if $d >= 0.5$, and small if $d >= 0.2$. To adjust these threshold values for the observed base rates, they are converted according to Eqn. 8, where $p1$ and $p2$ represent the proportions of groups described by the dichotomous variable, with $p_2 = 1 - p_1$:

$$threshold = \frac{d}{\sqrt{d^2 + \frac{1}{p_1, p_2}}} \quad (8)$$

[19] http://github.com/pytorch/fairseq

The adjusted effect size interpretation thresholds for WSD error correlation values as given in Table 3 are provided in Table 7. Adjusted thresholds for attack success correlations as given in Table 5 are summarized in Table 8.

A.6 Annotator instructions

The judges were presented with the following instructions for the described annotation tasks:

Your first task is to judge whether the meaning of the homograph as used in the given sentence is best described by the terms in the SENSE 1 cell or by those in the SENSE 2 cell. Please use the drop-down menu in the WHICH SENSE IS CORRECT? column to make your choice. If you think that neither sens captures the homograph's meaning, please select NONE from the options in the drop-down menu. If you think that the homograph as used in the given sentence can be equally interpreted both as SENSE 1 or SENSE 2, please select BOTH.

We're also asking you to give us your subjective judgment whether the sentence you've been evaluating makes sense to you, i.e. whether it's grammatical, whether it can be easily understood, and whether it sounds acceptable to you as a whole. Typos and spelling mistakes, on the other hand, can be ignored. Specifically, we would like you to assign each sentence a naturalness score, ranging from 1 to 5, according to the following scale:

- *1 = Completely unnatural (i.e. sentence is clearly ungrammatical, highly implausible, or meaningless / incoherent)*
- *2 = Somewhat unnatural (i.e. sentence is not outright incoherent, but sounds very strange)*
- *3 = Unsure (i.e. sentence is difficult to judge either way)*
- *4 = Mostly natural (i.e. sentence sounds good for the most part)*
- *5 = Completely natural (i.e. a well-formed English sentence)*

For instance a sentence like "John ate ten pancakes for breakfast." may get a ranking between 4 and 5, as it satisfies all of the above criteria. A sentence like "John ate green pancakes for breakfast." is grammatical but somewhat unusual and may therefore get a score between 3 and 4. "John ate late pancakes for breakfast.", on the other hand, does not sound very natural since pancakes cannot be "late" and may therefore be rated as 1 or 2. For this judgment we ask you to pay special attention to words in the neighborhood of the homograph. To submit your judgment please select the appropriate score from the drop-down menu in the DOES THE SENTENCE MAKE SENSE? column.

A.7 Examples of successful adversarial samples

Tables 13 - 18 list examples of successful adversarial attacks across the examined model architectures and dataset domains. As done throughout the paper, homographs are highlighted in blue, whereas the introduced attractors are emphasized in red.

Model	small	medium	large
OS18 Transformer	0.0542	0.1344	0.2121
OS18 LSTM	0.0666	0.1647	0.2581
OS18 ConvS2S	0.0710	0.1753	0.2740
WMT19 Transformer	0.0381	0.0949	0.1508
WMT19 LSTM	0.0458	0.1138	0.1803
WMT19 ConvS2S	0.0502	0.1247	0.1971

Table 7: Base-rate adjusted thresholds for the interpretation of WSD error prediction correlations.

Model	small	medium	large
OS18 Transformer	0.0339	0.0846	0.1345
OS18 LSTM	0.0338	0.0842	0.1340
OS18 ConvS2S	0.0328	0.0817	0.1301
WMT19 Transformer	0.0166	0.0414	0.0661
WMT19 LSTM	0.0178	0.0446	0.0712
WMT19 ConvS2S	0.0219	0.0548	0.0874

Table 8: Base-rate adjusted thresholds for the interpretation of attack success correlations.

Statistic	train	dev	test	held-out
# sentences	14,993,062	10,000	10,000	3,751,765
# words (EN)	106,873,835	71,719	71,332	26,763,351
# words/sentence (EN)	7.13	7.17	7.13	7.13
# words (DE)	100,248,893	67,185	66,799	25,094,166
# words/sentence (DE)	6.69	6,71	6.68	6.69

Table 9: Corpus statistics for the OS18 domain.

Statistic	train	dev (test18)	test14	test19	held-out
# sentences	4,861,743	2,998	3,003	1,997	1,215,435
# words (EN)	100,271,426	58,628	59,325	42034	25,057,036
# words/sentence (EN)	20.62	19.56	19.76	21.05	20.62
# words (DE)	93,900,343	54,933	54,865	42,087	23,467,086
# words/sentence (DE)	19.31	18.32	18.27	21.08	19.31

Table 10: Corpus statistics for the WMT19 domain.

Homograph	Sense 1	Sense 2	Sense 3
bat	*Chiroptera*, *Fledertier*, *Handflügler*, *Fledermaus*, *Flattertier*	*Schlagstock*, *Schlagholz*, *Baseballschläger*, *Baseballkeule*, *Schläger*	-
case	*Karton*, *Kiste*, *Päckchen*, *Packung*, *Schachtel*, *Kasten*, *Behälter*, *Box*	*Fall*, *Zustand*, *Sache*, *Gegebenheit*, *Lage*, *Kontext*, *Umstand*, *Status*, *Sachverhalt*, *Stand*, *Situation*	*Prozess*, *Gerichtsverfahren*, *Fall*, *Gerichtsverhandlung*, *Sache*, *Prozeß*, *Rechtsstreit*, *Ermittlung*, *Antrag*, *Rechtsfall*, *Gerichtsfall*, *Klage*, *Verhör*, *Rechtssache*
letter	*Sendschreiben*, *Papierbrief*, *Musterbrief*, *Anschreiben*, *Post*, *Schreiben*, *Brief*	*Buchstabe*, *Großbuchstabe*, *Charakter*, *Letter*, *Kleinbuchstabe*, *Zeichen*	-
spring	*Ringfeder*, *Spiralfeder*, *Sprungfeder*, *Feder*, *Tellerfeder*, *Federung*, *Gummifeder*	*Frühling*, *Lenz*, *Frühjahr*	*Quelle*, *Brunnen*, *Quell*, *Wasserquelle*
vacuum	*Vakuum*, *Nichts*, *Unterdruck*, *Leerraum*, *Leere*, *Luftleere*	*Industriestaubsauger*, *Staubsauger*, *Handstaubsauger*, *Teppichkehrer*, *Bodenstaubsauger*, *Allessauger*, *Sauger*, *Kesselsauger*	-

Table 11: Non-exhaustive examples of homograph-specific sense clusters.

Parameter	Transformer	LSTM	ConvS2S
batch size (subwords)	24,576	4,096	4,096
# total updates	100,000	600,000	600,000
# warm-up updates	4,000	-	-
# updates between checkpoints	1,000	4,000	4,000
# epochs between validations	1	1	1
optimizer	Adam	Adam	Adam
Adam betas	0.9, 0.98	0.9, 0.98	0.9, 0.98
learning rate	scheduled (*inverse_sqrt*)	0.0002 (+ decay)	0.0003 (+ decay)
# total parameters (OS18)	60,641,280	59,819,008	64,548,328
# total parameters (WMT19)	61,714,432	60,892,160	66,696,728
embedding size	512	512	512
Tied embeddings	Yes	Yes	Yes
hidden size	2,048	512	512
# encoder layers	6	5 (bidirectional)	8
# decoder layers	6	5	8
kernel size	-	-	3
dropout	0.1	0.2	0.2
label smoothing	0.1	0.1	0.1

Table 12: Training settings and model hyperparameters.

Source input / **O**riginal output / **P**erturbed output	Seed sense	Adv. sense
S: The Penguin was beating him with an (**old**) **bat**, but it was Gordon that pulled the trigger. **O**: Der Pinguin hat ihn mit einem **Schläger**✓ geschlagen, aber Gordon hat abgedrückt. **P**: Der Pinguin hat ihn mit einer **alten** **Fledermaus**✗ geschlagen , aber Gordon hat abgedrückt.	*club*	*animal*
S: I'm not going to relax until that thing its back in its (**simple**) **case**. **O**: Ich werde mich nicht entspannen, bis dieses Ding nicht seinen Rücken in seinem **Koffer**✓ hat. **P**: Ich werde mich nicht entspannen, bis das Ding nicht seinen Rücken in seinem **einfachen** **Fall**✗ hat.	*container*	*instance*
S: "They rest in their mother's (**hot**) **lap**, enjoying the ultimate bliss" **O**: "Sie ruhen im **Schoß**✓ ihrer Mutter, genießen das ultimative Glück" **P**: "Sie ruhen in der **heißen** **Runde**✗ ihrer Mutter, genießen das ultimative Glück"	*body part*	*circuit*
S: That's mighty neighbourly, but I got to play the (**big**) **organ** for the parson tonight. **O**: Das ist mächtig nachbarschaftlich, aber ich muss heute Abend **Orgel**✓ für den Pfarrer spielen. **P**: Das ist mächtig nachbarschaftlich, aber ich muss heute Abend das **Organ**✗ für den Pfarrer spielen.	*instrument*	*body part*
S: I'm just gonna write a (**high**) **note**, and then we'll go. **O**: Ich schreibe nur einen **Zettel**✓ und dann gehen wir. **P**: Ich schreibe einen **hohen** **Ton**✗ und dann gehen wir.	*writing*	*tone*

Table 13: Additional examples of successful attacks on the OS18 transformer. Homographs are blue, attractors are red.

Source input / Original output / Perturbed output	Seed sense	Adv. sense
S: I only sell (**good**) **arms** to people who fight clean wars! sure! **O**: Ich verkaufe nur **Waffen**✓ an Leute, die saubere Kriege bekämpfen. **P**: Ich verkaufe nur **gute Arme**✗ an Leute, die saubere Kriege bekämpfen.	*weapon*	*body part*
S: We've heard they're trying to raise (**new**) **capital** to rebuild their armies. **O**: Wir haben gehört, sie wollen **Kapital**✓ sammeln, um ihre Armeen aufzubauen. **P**: Wir haben gehört, dass sie eine **neue Hauptstadt**✗ aufziehen wollen, um ihre Armeen aufzubauen.	*money*	*city*
S: Did you charge the Donellys for five (**closed**) **cases** of vodka? **O**: Haben Sie die Donellys für fünf **Kisten**✓ Wodka berechnet? **P**: Haben Sie die Donellys für fünf **geschlossene Fälle**✗ Wodka berechnet?	*container*	*court case*
S: All units, repeat. that is a battered yellow van, no (**separate**) **plates**. **O**: An alle Einheiten, das ist ein gegrillter gelben Van, keine **Nummernschilder**✓. **P**: An alle Einheiten, das ist ein gegrillter gelben Van, keine **getrennten Teller**✗.	*number plate*	*dish*
S: Um, (**old**) **seals** tell the truth , but a sea lion's always lyin'? **O**: **Robben**✓ sagen die Wahrheit, aber ein Seelöwen lügt immer ? **P**: **Alte Siegel**✗ sagen die Wahrheit, aber ein Seelöwen lügt immer?	*animal*	*emblem*

Table 14: Examples of successful attacks on the OS18 LSTM. Homographs are blue, attractors are red.

Source input / Original output / Perturbed output	Seed sense	Adv. sense
S: - Oh, well, keep the (**small**) **change** and have a drink on me. **O**: Behalten Sie den **Rest**✓ und trinken Sie auf mich. **P**: Oh, nun, behalte die **kleine Veränderung**✗ und trink einen auf mich.	*coins*	*development*
S: Do you know how that (**specific**) **date** went, by any chance? **O**: Wissen Sie, wie das **Date**✓ gelaufen ist? **P**: Wissen Sie, wie das **Datum**✗ gelaufen ist?	*meeting*	*calendar date*
S: Goal! (public address) An amazing last-minute third goal that takes Greenock into the (**strong**) **lead**. **O**: Ein erstaunliches drittes drittes Ziel, das Greenock in die **Führung**✓ führt. **P**: Ein erstaunliches drittes Ziel, das Greenock in die **starke Spur**✗ führt.	*first place*	*clue*
S: I mean, you seem like someone who plots out every (**fucking**) **move**. **O**: Ich meine, Sie scheinen jemand zu sein, der jeden **Schritt**✓ aussticht. **P**: Ich meine, Sie scheinen jemand zu sein, der jede **verdammte Bewegung**✗ ausschüttet.	*action*	*movement*
S: You know, if we get hungry, we eat some chips, have some (**crazy**) **punch** ... **O**: Weißt du, wenn wir hungrig werden, essen wir ein paar Chips, haben etwas **Punsch**✓ ... **P**: Weißt du, wenn wir hungrig werden, essen wir ein paar Chips, haben einen **verrückten Schlag**✗ ...	*drink*	*hit*

Table 15: Examples of successful attacks on the OS18 ConvS2S. Homographs are blue, attractors are red.

Source input / **O**riginal output / **P**erturbed output	**Seed sense**	**Adv. sense**
S: Copenhagen - Copenhagen, Denmark's (**financial**) **capital**, wants to be the world's first CO2-neutral city by 2025. **O**: Kopenhagen - Kopenhagen, die **Hauptstadt**✓ Dänemarks, will bis 2025 die erste CO2-neutrale Stadt der Welt sein. **P**: Kopenhagen - Kopenhagen, das **Finanzkapital**✗ Dänemarks, will bis 2025 die erste CO2-neutrale Stadt der Welt sein.	*city*	*money*
S: This is done by pricking the earlobe with a small lancet and taking a (**real**) **drop** of blood. **O**: Dies geschieht, indem der Ohrwurm mit einem kleinen Lancet geprickt wird und ein **Tropfen**✓ Blut eingenommen wird. **P**: Dies geschieht, indem der Ohrwurm mit einem kleinen Lancet geprickt wird und ein **richtiger** Blut**abfall**✗ entsteht.	*drop of liquid*	*decrease*
S: One (**small** ~~positive~~) **note** was from the Republic of Ireland, which saw its PMI grow to 57.3, its highest level since the end of 1999. **O**: Eine positive **Anmerkung**✓ war die aus der Republik Irland, wo das PMI auf 57,3 anstieg, das höchste Niveau seit Ende 1999. **P**: Ein **kleiner Schein**✗ stammt aus der Republik Irland, wo das PMI auf 57,3 anstieg, das höchste Niveau seit Ende 1999.	*remark*	*paper money*
S: His epoch-making (**full**) **record** "Free Jazz" was released by Atlantic Records at the dawn of that decade. **O**: Seine epochale **Platte**✓ "Free Jazz" wurde zu Beginn des Jahrzehnts von Atlantic Records veröffentlicht. **P**: Seine epochale **Aufzeichnung**✗ "Free Jazz" wurde zu Beginn des Jahrzehnts von Atlantic Records veröffentlicht.	*musical medium*	*document*
S: After winter delivered an early dose of (**natural**) **spring** last week, temperatures dropped again on Monday to a high of just 15.8C in the city. **O**: Nachdem der Winter vergangene Woche eine frühe **Frühjahr**sdosis✓ geliefert hatte, fielen die Temperaturen am Montag wieder auf einen Höchstwert von nur 15,8C in der Stadt. **P**: Nachdem der Winter letzte Woche eine frühe Dosis **Naturquelle**✗ lieferte, fielen die Temperaturen am Montag wieder auf einen Höchstwert von nur 15,8C in der Stadt.	*season*	*water source*

Table 16: Examples of successful attacks on the WMT19 transformer. Homographs are blue, attractors are red.

Source input / Original output / Perturbed output	Seed sense	Adv. sense
S: A Thousand Splendid Suns is a story of two women's lives in Afghanistan, where women are equal, as a table or the (**last**) **chair**. **O**: Ein Thousand Splendid Seine ist eine Geschichte von zwei Frauen in Afghanistan, wo Frauen gleich sind, als Tisch oder **Stuhl**✓. **P**: Ein Thousand Splendid Seine ist eine Geschichte von zwei Frauen in Afghanistan, wo Frauen gleich sind, als Tisch oder als **letzter Vorsitzender**✗.	*furniture*	*chairperson*
S: See a (**small** ~~rapid~~) **drop** in your CO level once you stop smoking. **O**: Sehen Sie sich einen schnellen **Rückgang**✓ Ihrer CO-Ebene an, sobald Sie das Rauchen einstellen. **P**: Sehen Sie einen **kleinen Tropfen**✗ auf Ihrem CO-Niveau, sobald Sie aufhören, Rauchen zu beenden.	*decrease*	*drop of liquid*
S: And moreover - each of our guests will get a (**different** ~~small~~) **present**! **O**: Und darüber hinaus wird jeder unserer Gäste ein kleines **Geschenk**✓ bekommen! **P**: Und darüber hinaus wird jeder unserer Gäste eine **andere Gegenwart**✗ bekommen!	*gift*	*current time*
S: A (**new**) **record** of every transaction made is kept, allowing for a complete audit if necessary. **O**: Ein **Datensatz**✓ jeder Transaktion wird gehalten, so dass erforderlichenfalls eine vollständige Prüfung möglich ist. **P**: Ein **neuer Rekord**✗ jeder Transaktion wird gehalten, so dass erforderlichenfalls eine vollständige Prüfung möglich ist.	*document*	*achievement*
S: Britain's new trade deals with non-EU countries would also probably involve (**political** ~~worse~~) **terms**. **O**: Die neuen Handelsvereinbarungen Großbritanniens mit Nicht-EU-Ländern würden wahrscheinlich auch schlechtere **Bedingungen**✓ beinhalten. **P**: Großbritanniens neue Handelsabkommen mit Nicht-EU-Ländern würden wahrscheinlich auch **politische Begriffe**✗ beinhalten.	*demand*	*expression*

Table 17: Examples of successful attacks on the WMT19 LSTM. Homographs are blue, attractors are red.

Source input / Original output / Perturbed output	Seed sense	Adv. sense
S: Not to mention (non) uniform loading and soring fingers, contaminated with (**common**) **lead**. **O**: Ganz zu schweigen von (nicht) einheitlichen Lade- und Sortierfingern, die mit **Blei**✓ kontaminiert sind. **P**: Ganz zu schweigen von (nicht) einheitlichen Lade- und Sortierfingern, die mit einer **gemeinsamen Führung**✗ kontaminiert sind.	*metal*	*first place*
S: If the symbol "&gt" is displayed, keep entering (**greek**) **letters** until predictive options are displayed. **O**: Wenn das Symbol "&gt" angezeigt wird, erhalten Sie die Eingabe von **Buchstaben**✓, bis prognostizierte Optionen angezeigt werden. **P**: Wenn das Symbol "&gt" angezeigt wird, erhalten Sie immer wieder Gruß**briefe**✗, bis prognostizierte Optionen angezeigt werden.	*character*	*message*
S: This film is not about dialogue or a (**little** ~~stringent~~) **plot**, but all about atmosphere - a feverish dream that has become a film. **O**: In diesem Film geht es nicht um einen Dialog oder um eine strenge **Handlung**✓, sondern um die Atmosphäre - ein feverser Traum, der zu einem Film geworden ist. **P**: In diesem Film geht es nicht um Dialog oder ein **wenig Grundstück**✗, sondern alles über die Atmosphäre - ein feverser Traum, der zu einem Film geworden ist.	*story*	*tract of land*
S: Manufacture of products from silicone and rubber, Production of springs, Manufacturing of springs, Winding of (**small**) **springs**. **O**: Herstellung von Produkten aus Silikon- und Gummi, Herstellung von Quellen, Herstellung von Quellen, **Federn**✓. **P**: Herstellung von Produkten aus Silikon- und Gummi, Herstellung von Quellen, Herstellung von Quellen, Winding von **kleinen Quellen**✗.	*device*	*water source*
S: In 1980, financial assets - (**large**) **stocks**, bonds, and bank deposits - totaled around 100% of GDP in the advanced economies. O; Im Jahr 1980 belief sich das Finanzvermögen - **Aktien**✓, Anleihen und Bankeinlagen - in den hochentwickelten Volkswirtschaften rund 100% des BIP. **P**: Im Jahr 1980 belief sich das Finanzvermögen - **große Bestände**✗, Anleihen und Bankeinlagen - in den hochentwickelten Volkswirtschaften rund 100% des BIP.	*investment*	*inventory*

Table 18: Examples of successful attacks on the WMT19 ConvS2S. Homographs are blue, attractors are red.